DOCUMENT RESUME

ED 359 047 SE 053 466

AUTHOR	Schoen, Harold L.
TITLE	Assessment Issues from a Mathematics Education Perspective. The Mathematics Outlook.
INSTITUTION	ERIC Clearinghouse for Science, Mathematics, and Environmental Education, Columbus, Ohio.
SPONS AGENCY	Office of Educational Research and Improvement (ED), Washington, DC.
PUB DATE	May 93
CONTRACT	RI88062006
NOTE	11p.
AVAILABLE FROM	ERIC Clearinghouse for Science, Mathematics, and Environmental Education, 1929 Kenny Road, Columbus, OH 43210-1080 ($1.50).
PUB TYPE	Information Analyses - ERIC Clearinghouse Products (071) -- Guides - Non-Classroom Use (055)
EDRS PRICE	MF01/PC01 Plus Postage.
DESCRIPTORS	Curriculum Development; Diagnostic Teaching; *Educational Change; *Educational Testing; Elementary Secondary Education; Equal Education; *Evaluation Methods; Informal Assessment; International Studies; *Mathematics Achievement; *Mathematics Curriculum; Mathematics Education; *Mathematics Instruction; Standardized Tests
IDENTIFIERS	National Assessment of Educational Progress; NCTM Curriculum and Evaluation Standards; Performance Based Evaluation

ABSTRACT

Recent reports by the National Council of Teachers of Mathematics (NCTM), the National Research Council, the Mathematical Sciences Education Board, and the Board on Mathematical Sciences emphasize that coordinated improvement in the content of mathematics curriculum, classroom instruction, and teacher education is the key to improving mathematics education. Consequentially, methods of assessing students' mathematics achievement also need to be reevaluated. This bulletin focuses on the impact that assessment has and can have on mathematics curriculum and instruction. The bulletin is divided into two sections. The first section addresses assessment done by the teacher as part of classroom instruction. Topics discussed include: teacher competencies in assessment, assessment techniques as recommended by the NCTM, reform of large-scale mathematics testing, and the need for equal opportunities for all students to learn mathematics. The second section discusses the impact that certain uses of large-scale testing can have on curriculum revision efforts. Topics discussed include: the perception of testing as a gatekeeper to curriculum reform and the revision of standardized tests to reflect current curricular reforms. A list of 36 references is included. (MDH)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED359047

The Mathematics OutLook

ERIC CSMEE
Informational Bulletin

Assessment Issues From A Mathematics Education Perspective

by: Harold L. Schoen

Clearinghouse for Science, Mathematics, and Environmental Education — *May 1993*

By the mid 1980s, mathematics education in the United States was under intense pressure to make major changes at all levels. Many students were dropping out of mathematics at their earliest opportunities, and others were being held for an inordinate amount of time in routine arithmetic skill courses. African-Americans and Hispanics were most likely to be in these categories. Undergraduate students majoring in mathematics-related fields were decreasing in number, and increasing numbers of students, especially at the graduate level, were not U.S. citizens. Business and industry were decrying the inability of new employees to deal with mathematical problems or mathematical content beyond the most basic level. This problem was corroborated by the National Assessment of Educational Progress, which showed that a nationally representative sample of students could do basic computation but was unable to apply mathematics in real world situations. School and university mathematics were being criticized on other fronts for being outdated and for failing to prepare students to deal with the sophisticated technology they would encounter in the workplace. Finally, a barrage of criticism and calls for change followed the announcement that American students in grades eight and eleven scored well below the median in the Second International Mathematics Study. This performance was much lower than that of students in highly industrialized Japan and Western European nations.

In response to these pressures, the National Council of Teachers of Mathematics organized its Standards Project in 1986, which has now resulted in the publication of two sets of recommendations that together provide a coordinated vision of the school mathematics curriculum, the teaching of mathematics, and the assessment of mathematical abilities (NCTM, 1989, 1991). Furthermore, the mathematics and mathematics education communities through publications of professional organizations have now articulated their collective professional judgment of what exemplary practice and recent research (in teaching and learning as well as in mathematics itself) suggest for the mathematics curriculum. Closely coordinated with the NCTM standards are recommendations for reforming the undergraduate mathematics curriculum, produced jointly by the National Research Council (1991), the Mathematical Sciences Education Board, the Committee on the Mathematical Sciences in the Year 2000, and the Board on Mathematical Sciences. Recommendations for the mathematical content for prospective teachers have also been made (Mathematical Association of America, 1991).

These reports all emphasize that coordinated, cross-level improvement in the content of the curriculum, in classroom instruction, and in teacher education is the key to improving mathematics education. In addition to curriculum, instruction, and teacher education, methods of assessing the mathematical performance of students must be reevaluated in light of the new curricular goals. The above sets of recommendations are in agreement on this last point, too, but only the two volumes from the NCTM address assessment in a substantive way (NCTM, 1989, 1991).

The term assessment is meant to have a broader meaning than testing in traditional ways, most notably the ubiquitous paper-and-pencil *objective* test. Assessment is not necessarily restricted to formal evaluations of student achievement, instruction, or program. This paper focuses on assessment as it impacts, or could impact, mathematics curriculum and instruction. It is divided into two main sections: The first addresses assessment done by the teacher as part of classroom instruction. The second section discusses the impact that certain uses of large-scale testing, prepared by measurement specialists outside the classroom, can have on curriculum revision efforts.

Assessment in Support of Instruction

Teachers face daily unique problematic situations in their classrooms. At any time, a student may illustrate confusion, yet a partial understanding of a concept, in a way that is entirely new to the teacher. Because these situations fall outside any existing theory, the teacher will have no source of directly applicable rules available. In order to respond competently, the teacher must improvise and test strategies of his or her own design in the situation. In short, the teacher must be a reflective practitioner, applying in ad hoc ways all experience, pedagogical knowledge of content, knowledge of students, and the interaction of these components (Schon, 1990). A reflective, highly competent teacher will be almost continuously assessing students' understanding, their motivation to work on certain tasks, their readiness to proceed to new activities, their ability to work together, the effectiveness of a lesson plan as he or she implements it, and so on. Assessment, in this broad sense, is a very important part of instruction.

Professionals other than teachers, such as medical doctors, lawyers, and clinical psychologists, generally conduct assessments and make diagnoses that are required in their work. These professionals may employ technicians and testing instruments when they judge it to be appropriate, but they reserve the right to analyze the assessment results and prescribe a

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)
☒ This document has been reproduced as received from the person or organization originating it.
☐ Minor changes have been made to improve reproduction quality.
• Points of view or opinions stated in this document do not necessarily represent official

BEST COPY AVAILABLE

SE 053 464

strategy or treatment that combines their best understanding of all pertinent aspects of a situation. The ability to conduct assessments in their domain of expertise and to translate the results into directions for practice is a major part of what characterizes a professional in that domain.

The teacher, as the professional in the classroom, is in the best position to understand all pertinent variables that may help solve local problems of instruction and learning. No one else has experience in this classroom, understanding of how these students learn and what motivates them, or understanding of how to teach the subject matter to these students, what is likely to be difficult for them, and what content came before and what comes next in the curriculum. For the most part, however, teachers have not been well educated in assessment techniques and strategies. The education of teachers is usually far more prescriptive than the education of other professionals, focusing on rules about how to handle classroom disruptions and very specific teaching activities that can be inserted directly into lesson plans. Strategies for assessing students' understanding of mathematics (or any other subject matter) are usually given little attention beyond a brief unit on interpreting results of norm-referenced tests and some experience writing and scoring tests during student teaching (NCTM, 1991; Wolf, Bixby, Glenn, & Gardner, 1991). As a result, teachers are much better at managing a class than they are at individual or group assessment, diagnosis, and prescription. Assessing, to most of them, means giving paper-and-pencil tests, either large-scale tests developed by measurement specialists or their own tests for purposes of grading.

Recognizing the need for teachers to become more skilled at assessment, a joint committee of the American Federation of Teachers, the National Council of Measurement in Education, and the National Education Association developed the following seven standards for teacher competence in educational assessment (AFT, NCME & NEA, 1990). Teachers should be skilled in:

1. choosing assessment methods appropriate for instructional decisions;
2. developing assessment methods appropriate for instructional decisions;
3. administering, scoring, and interpreting the results of both externally produced and teacher-produced assessment methods;
4. using assessment results when making decisions about individual students, planning teaching, developing curriculum, and school improvement;
5. developing valid pupil grading procedures that use pupil assessments;
6. communicating assessment results to students, parents, other lay audiences, and other educators; and
7. recognizing unethical, illegal, and otherwise inappropriate assessment methods and uses of assessment information.

The above standards are certainly important for all teachers. These standards, especially the first four, assume, however, that teachers possess what Shulman (1986) called *pedagogical content knowledge*. As a prerequisite to the above competencies, teachers need a view of what is important for students to understand and be able to do in the subject matter they are teaching, which student behavior indicates an understanding of particular concepts in the subject matter, difficulties that their students are likely to encounter in trying to learn particular topics, how ideas in the curriculum are connected to one another, and multiple ways to represent concepts that will better focus the assessment on deficiencies in an individual student's understanding.

In fact, Shulman (1991) argued that the study of effective teaching in a generic sense has limited potential. Rather, teaching can only be understood by studying the knowledge of pedagogy that is required to effectively teach specific content to specific students in specific contexts. The same sort of limitation (viewing assessment as content-free and context-free) seems to apply to a generic set of assessment competencies like the AFT-NCME-NEA standards given earlier.

As an illustration for a mathematics teacher, consider a junior high pre-algebra student who has no difficulty with equations like the following:

$$x + 3 = 10$$

$$2x - 8 = 15$$

The student has persistent difficulty, however, when the sides of the equations are switched despite the teacher's best efforts to point out the error. Frustrated teachers in such situations often respond by blaming the student's carelessness or earlier teachers' incompetence or lack of thoroughness. Such unproductive responses are less likely if the teacher knows that this is a very common error pattern due largely to the different uses of the equal sign in arithmetic and in algebra. In arithmetic, the equal sign is almost always a signal to *do something* to whatever is on the left side of the equation to *get the answer* that is then placed on the right side. In algebra, however, the same symbol means the numerical value of the left side is the same as that of the right side. Six or eight years of a student's successfully viewing a concept like *equal* in one way cannot be undone easily or quickly. Herscovics and Kieran (1980) have designed and validated an instructional sequence that begins with the arithmetic meaning of equality and leads the student to expand that meaning to the algebraic concept.

Researchers have analyzed how students at different levels learn other concepts and topics in mathematics and have identified the kinds of errors and misconceptions that are likely. For example, a great deal of research has focused on elementary and middle school topics like fractions, decimals, and story problems, which have different mathematical structures (e.g., Hiebert & Behr, 1988). Algebraic or pre-algebraic topics (like equality in the earlier example), variables, and graphs have also been studied in detail from this perspective (cf. Wagner & Kieran, 1989; Leinhardt, Zaslavsky, & Stein, 1990). Instructional sequences that help students learn these concepts and avert or correct misconceptions have also be developed and validated for many concepts (cf. Trafton, 1989; Coxford, 1988). Teachers who are not well versed in these pedagogical aspects of mathematics will be limited in their ability to assess a student's performance and to make appropriate adjustments in instruction.

For the vision of the mathematics curriculum common to these sets of recommendations to become a reality, textbooks and other instructional material that support these curricular goals will need to be developed. Assessment techniques and

assessment instruments should also support the curriculum and instruction. Thus, all assessment, whether developed by the teacher or by persons or agencies outside the classroom, should be aligned with the content topics of the curriculum and should reinforce the view of what is important in the curriculum and in classroom instruction.

Assessment Techniques

The NCTM recommends that teachers should be able to do the following:

1. use a variety of assessment methods to determine students' understanding of mathematics;
2. match assessment methods with the developmental level, the mathematical maturity, and the cultural background of the student;
3. align assessment methods with what is taught and how it is taught;
4. analyze individual students' understanding of, and disposition to do, mathematics so that information about their mathematical development can be provided to the students, their parents, and pertinent school personnel; and
5. base instruction on information obtained from assessing students' understanding of, and disposition to do, mathematics (NCTM, 1991, p. 110).

To be able to do all of this, teachers must surely have a solid pedagogical knowledge of mathematics, as described earlier in this paper, at least at the level that they are teaching, and a clear understanding of how students learn. Like teaching, assessing as a part of instruction cannot be meaningfully separated from the teacher's knowledge of the subject matter and understanding of how students learn that subject matter. In addition to the teacher's knowledge of the subject matter and of the students, the teacher's repertoire should include a variety of assessment techniques and a plan that ties assessment to its instructional purpose. A number of instructional purposes that might drive the choice of an assessment method are given in Table 1 (p.4), which is a brief version of a table found in NCTM (1989; pp. 200-201).

Few of the assessment methods in Table 1 are new. They reflect and reemphasize, however, the rich variety of student outcomes that should be intended in a mathematics class. High scores on a standardized multiple-choice or short-answer test are a desirable outcome but by no means the only one. Since the assessment method gives a strong message to the student about what is important (Lester & Kroll, 1990), all desirable instructional outcomes should be assessed. Stated differently, if an instructional outcome is not assessed, students are likely to consider it to be unimportant and not worth their time and attention, despite the teacher's emphasis or exhortations to the contrary.

If NCTM recommendations such as solving applied problems, reasoning mathematically, communicating mathematical ideas using tools such as calculators, and working with other students are desired instructional outcomes, then some of the assessment that "counts" should include observation and analysis of students doing mathematics under these conditions. For example, students could be given the following problem to work on in a small group (California Mathematics Council, 1989; p. 21).

> We have reached into this bag of blocks 6 times and have pulled out 3 red blocks, 1 green block, and 2 blue blocks. If you reached into the bag and pulled out another block, what color do you think it would be? Explain why you think it would be that color. How could you get more information?

Questions concerning the students' problem-solving performance that the teacher as observer might want to use as a focus are the following:

> Do students have a systematic way of organizing and recording information?
>
> Do they relate a problem to other similar problems?
>
> Are they able to express their ideas orally?
>
> Are they able to come up with ideas for getting more information?

On the other hand, if the teacher wanted to assess the students' disposition toward doing mathematics, the focus might be:

> Do students plan before acting and revise their plan as necessary?
>
> Do they stick to the task without being easily distracted?
>
> Do they use supplementary tools such as calculators as needed?
>
> Do they support their arguments with evidence?
>
> Do they complete the task?
>
> Do they review their solution process and their result?

Similarly, if the teacher was interested in how well students worked as a group, he or she would focus on group interaction and communication behaviors such as the following:

> Do students engage in discussions in order to clarify and communicate their ideas to others?
>
> Do they describe their problem-solving processes clearly enough so that they are replicable?
>
> Do they have the confidence to make a report to the whole class?
>
> Do they capably and fairly represent a group consensus?
>
> Do they synthesize and summarize individual and group thinking?

Mathematics educators are also beginning to borrow the idea of portfolio assessment from art and writing teachers. One type of portfolio that shows a student's development would contain a sampling of a student's earlier and more recent work

Table 1. Purposes and Methods of Assessment (NCTM, 1989; pp.200-201)

Purposes (examples of questions asked)	Assessment Methods

Diagnostic

Purposes (examples of questions asked):
- What does this student understand about the concept or procedure?
- What aspects of problem solving are causing difficulty?
- What accounts for this student's unwillingness to attempt new problems or to see the application of previously learned materials?

Assessment Methods:
- Oral questions that ask students to explain their procedure
- Focused written tasks
- Directed test items
- Observation

Instructional Feedback

Purposes:
- What do students know about the material presented?
- Can students apply their learning to new situations?
- Do students understand the connections among ideas?
- How shall I pace instruction?
- Does the class need more intensive review or more challenging material?

Assessment Methods:
- Written tests, including those that require differential methods for solutions to problems
- Class presentations
- Extended problem-solving projects
- Observation of class
- Take-home tests
- Homework, journals
- Group work and projects

Grading

Purposes:
- How well has this student understood and integrated the material?
- Can this student apply his or her learning in other contexts?
- How prepared is this student to proceed to the next grade or level?

Assessment Methods:
- Extended problem solving projects
- Papers or written arguments that demand thoughtful inquiry about a mathematical topic
- Written tests that present problems with a range of difficulty based on expectations for the course
- Oral presentations

Generalized mathematical achievement

Purposes:
- How does the general mathematical ability of this student compare with others or with a national norm?

Assessment Methods:
- Standardized achievement tests

Program evaluation

Purposes:
- How effective is this instructional program in achieving our goals for mathematical learning?

Assessment Methods:
- Student interviews
- Performance tests
- Criterion-referenced tests
- Observation of class discussions
- Success of students who have completed the program

that illustrates various kinds of mathematical performances (Wolf, et al., 1991). Such a portfolio in mathematics might include a student's written descriptions of the results of practical or purely mathematical projects, extended analysis of problem situations and investigations, descriptions and diagrams of problem-solving processes, statistical studies and graphic representations of data, responses to open-ended questions or homework problems, copies of awards or prizes, video, audio, or computer-generated examples of student work, a mathematical biography updated annually, and excerpts from the student's mathematical journal (California Mathematics Council, 1989). The result of keeping portfolios is that teachers, students, and parents have access to a continuous body of work that is an indication of the student's cognitive and affective development over time.

Other assessment techniques and ways for teachers to record and analyze data resulting from using these techniques are described in many sources (e.g., California Mathematics Council, 1990; Clarke, Clarke, & Lovitt, 1990; Lester & Kroll, 1990, 1991; NCTM, 1989, 1991; Webb & Briars, 1990). More extensive guidelines for assessment in mathematics classrooms will be appearing soon. For example, the 1993 Yearbook of the NCTM is tentatively entitled *Assessment in the Mathematics Classroom*, and NCTM is presently working with an author team on a publication for teachers with the working title, *Mathematics Assessment: Myths, Models, Good Questions, and Practical Suggestions*.

Large-scale Mathematics Testing

Large-scale paper-and-pencil testing, typically in a multiple-choice format, has been highly visible in American education for most of this century. The results of these tests usually have little to say to an individual teacher about how to improve classroom instruction. Rather, they serve to order students and to compare them to national norms in the case of standardized tests, to give an interpretation of student performance in broad mathematical content areas in the case of the National Assessment of Educational Progress and most state assessments, or to give a country-by-country comparison of average mathematics achievement. Most people agree that these tests serve their purposes quite well.

The present nationwide furor over educational testing arises from many misuses of the tests. For example, they are often used to place students into different academic tracks; to judge the quality of a curriculum, teachers, schools, and the U. S. educational system as a whole; and to "drive" curriculum and instruction (Popham, 1987). In the extreme, standardized tests have been used for such high-stakes purposes as determining the funding level for a district or school, the salary of individual teachers, and whether individual students will graduate (Smith, 1991).

Large-scale achievement tests are, in theory, measurement instruments that unobtrusively and reliably quantify the more or less stable achievement of the student, "as if the student who ticks off items were inert matter to be assayed and as if all the agency and inquiry belongs to those doing the measuring" (Wolf, et al., 1991, p. 46). If in fact as well as in theory such tests were unobtrusive and had no significant effect on students, teachers, curriculum, or instruction, they would probably be of little concern to the mathematics education profession. However, a growing body of recent research is beginning to describe a substantial impact of misuses of large-scale tests on curriculum and instruction, as well as on students and teachers themselves. In a review of this research, Wolf, et al. (1991) summarize the effects of this technically elegant educational measurement system as follows:

> It distorts instruction, underscores inequities in access to education, and forecloses on students and teachers becoming active participants in signal debates over the standards that will be applied to their work. (p. 32)

Primarily as a result of such research findings, there is growing recognition among testing experts, educators, and political leaders that our large-scale testing practices need to be reformed (e.g., National Commission on Testing and Public Policy, 1990).

Joining in this general concern about testing practices, many leaders in mathematics and mathematics education see the widespread and often ill-advised uses of standardized testing as a major barrier to the success of the present mathematics curriculum improvement effort (e.g., Kulm, 1990). The primary reason for this concern is that the distortions of instruction alluded to by Wolf et al. are in direct conflict with the goals of the mathematics curriculum improvement effort. The distortions include narrowing of instructional focus; fragmentation of content; emphasis on isolated individual efforts by students with no tools beyond paper and pencil; use of basic skill mastery as a gatekeeper to more interesting mathematics; and de-emphasis on reasoning, problem solving, and communication.

Mathematics for All Students

A very important goal of the mathematics reform effort is to provide all students with equal opportunities to learn mathematics beyond arithmetic skills. A barrier to that goal is the use of test scores as gatekeepers to mathematics courses, which, to make matters worse, also results in a disproportionate exclusion of African-American and Hispanic students from substantial mathematics experiences (Oakes, 1990; Wolf, et al., 1991). Related to this last concern is another artifact of our testing system that is troubling to mathematics educators. Wolf et al. (1991) describe it as follows:

> In essence, for all the sophistication of our testing system, the concern for ranking and classifying has led to the acceptance of a significant proportion of failure or poor performance as "natural." The attention to...relative information overshadows the responsibility to see that all students learn and the necessity to provide explicit information about students' current levels of achievement. (p. 44)

The acceptance in the United States of a significant proportion of failure or poor performance is nowhere more evident than in mathematics. Many Americans are not ashamed to admit publicly that they could never do mathematics and to justify that inability by a lack of "a math gene" that somehow gives a mysterious power to a select few—a power that is inescapably beyond the grasp of most (NRC, 1989). This tendency to

accept failure in mathematics as natural is dramatically illustrated in the series of cross-cultural studies of mathematics achievement in the United States, Japan, and Taiwan (Stevenson, Lummis, Lee, & Stigler, 1990; Stigler, Lee, & Stevenson, 1990).

Focusing on first-grade and fifth-grade classrooms (40 in each country) in similar schools in the three countries, these studies included testing mathematics achievement, observing classrooms, and measuring students', teachers', and mothers' attitudes and attributions about mathematics. Overall, the superiority of the mathematical knowledge of the Asian students suggested by the Second International Mathematics Study (McKnight, et al., 1987) was found to be strong and consistent across various kinds of mathematical knowledge (Stigler, et al., 1990). The Asian students not only displayed superior skill in computation but were even more impressive in their performance on tasks that required an understanding of the structure of mathematics. American students, in contrast, tended to approach problems that required some understanding or reasoning in a routinized manner, typically doing without much thought some sort of calculations on all numbers in the problem. American students also had much more difficulty relating their knowledge of mathematics to the real world.

Despite this markedly inferior mathematical performance of their children, American mothers expressed a higher level of satisfaction than did Asian mothers with both their children's performance in mathematics and the mathematical education they were receiving in school:

> In the educational philosophies of Taiwan and Japan, nearly all children are believed to be capable of understanding the content of the elementary school curriculum. Lack of achievement is attributed to a failure to work hard.
>
> Americans, in contrast tend to place more emphasis on innate ability as a major reason for variations in achievement. In a sincere effort to provide experiences that children at different levels of ability can manage, American children are divided into groups or tracks according to their presumed ability levels.
>
> When adults convey an impression that some children are not expected to keep up with others, the children's motivation is likely to be diminished. Thus, paradoxically, well-meant allowances for different levels of ability may actually run the risk of decreasing the motivation to learn and undermining the achievement of many American children. (Stevenson, et al., 1990, p. 32)

This final observation by the authors is corroborated by a large body of research on ability grouping (e.g., Oakes, 1990).

From a different perspective, there is an interesting probability argument against taking very seriously test scores and other a priori criteria as gatekeepers to mathematics courses. The argument is a variation on one put forth by Paulos (1988) in his delightful book, *Innumeracy*. He pointed out that under certain reasonable assumptions only about 20% of people identified as having cancer by a 95%-accurate medical test can actually be expected to be correctly identified.

Suppose a test (perhaps in combination with other criteria) is being used to place students into, say, general mathematics rather than algebra. No criteria can be perfectly reliable and take into account all salient variables (for example, how hard the student will work during the coming year) that may contribute to a student's success or failure in algebra. An assumption of some random error, say 80% accuracy, in placement criteria leads to a rather disturbing result. For purposes of illustration, suppose that there is a theoretically "correct placement" for each of 1000 students, namely, the correct placement for 900 of the 1000 in algebra. How accurately will the students be placed into these courses? Eighty of the 100 students who should be in general mathematics will be accurately placed on average, but 180 (20%) of the 900 students who should be in algebra will be incorrectly placed in general mathematics. If these criteria are rigidly applied, *260 students will be placed in general mathematics, and 180 (nearly 70%) of them should be in algebra.* If there are any prejudices against low socioeconomic or under-represented minority students built into the placement criteria, and research suggests that often there are such prejudices (Oakes, 1990), the mistaken exclusion of students in these groups from algebra will be even more dramatic.

In summary, the evidence concerning the negative effects of ability grouping in mathematics, especially the exclusion of traditionally under-represented minorities, is overwhelming (e.g., Oakes, 1990). The fact that this practice persists in the face of all the evidence and the moral arguments against it is a dramatic illustration of the strong beliefs in this country that standardized test scores are a valid measurement of a person's natural ability or potential for success.

Testing as a Gatekeeper to Curriculum Reform

A broader reason for mathematics educators' concern about large-scale testing is that, when nationally publicized, student performance on these tests can have very serious ramifications for the overall direction of the mathematics curriculum. Even if it is not designed to measure curriculum outcomes, large-scale testing is a very powerful political vehicle for curriculum reformers and their opponents alike, who recognize the power of the press and the public's deeply ingrained "bottom line, win-or-lose" achievement test view of what constitutes success in an educational program. Low or declining scores on such tests can help to slow or virtually stop a curriculum improvement movement, as the widely publicized standardized test score declines in the early 1970s did for the "modern mathematics" efforts (National Advisory Committee on Mathematical Education, 1975). Poor results on large-scale tests, on the other hand, may provide some of the impetus for reforming the curriculum, as the relatively poor performance of American students in the Second International Mathematics Study has done for the present reform efforts (McKnight, Crosswhite, Dossey, Kifer, Swafford, Travers, & Cooney, 1987; NRC, 1989).

A good example of this phenomenon can be seen in some of the publicity about the recent statewide comparisons of mathematics achievement conducted by the National Assessment of Educational Progress (NAEP). In an article in *Newsweek*, June 17, 1991, entitled "A Dismal Report Card," Bill Honig,

California's superintendent of schools, used the relatively poor results for his state as an opportunity to criticize local schools for continuing to resist reforms like group problem solving and creative thinking. In his words, "It's like we have a cure for polio, but we're not giving the inoculation" (p. 64-65). Shirley Hill, chairperson of the Mathematical Sciences Education Board was quoted in the same article, using the NAEP results to argue for reform: "Until recently, the public was perfectly happy with students who could do the basics of adding and subtracting. Now we realize how much more students need to know, and people are going to be upset that they don't know it" (p. 67).

On the other hand, these same NAEP results were an occasion for others to argue against reforms like the NCTM standards. A June 7, 1991, article in *The Wall Street Journal* stated:

> There isn't any proof in this data that the recommended NCTM practices are clearly more effective than any other practices," said Archie LaPointe, executive director of NAEP. John Saxon, a conservative critic of the NCTM standards, was more blunt. "I don't believe the states that did well are chasing (NCTM's) elusive star. (p. B1)

The authors of the same article go on to note that "students in high-scoring states reported spending more classroom time than others working on problems directly from math textbooks, the kind of do-alone, routine activity that the math-teaching group wants de-emphasized" (p. B1).

In reality, the NAEP results provide no evidence concerning the effectiveness of the NCTM standards and no direct evidence of the need for any particular educational reform, although unacceptably low scores on NAEP suggest the need to change something. The NCTM standards appeared in 1989 and, when NAEP was administered, most teachers in the country had probably not even heard of them, let alone used them to guide the curriculum or instruction in their classrooms. Furthermore, NAEP made no claims to be in alignment with the standards, and even if it had been, the NAEP results could hardly be attributed to the mathematics curriculum. In fact, NAEP's state rankings were very similar to those that have resulted year after year when other achievement scores such as SAT scores have been compared. It makes little sense to now attribute those same rankings to state-by-state differences in the level of implementation of the NCTM standards.

For mathematics educators interested in seeing major curriculum revisions have an impact and get a fair trial in American schools, this sort of misuse of tests and test results for purposes of judging in the press the success of a new curriculum is of great concern. Such misguided "curriculum evaluation" surely contributes to the cycle of one failed educational reform effort after another in this country. One can always argue persuasively to the general public with these "hard scientific data" from nationally representative samples of students that whatever is now going on in curriculum and instruction is a failure, and, therefore, a reform is justified. But just as surely, the bottom-line, large-scale test score criterion for success of a curriculum dooms the reform to failure once it becomes viewed as "the present curriculum," because test scores will never be high enough to please everyone.

Clearly, if education is to break out of this vicious cycle, our win-or-lose, testing mentality must change. Society has a moral obligation to care for and educate its children, and, in particular, to give them the best mathematics education of which we are capable. Scores on appropriate, large-scale tests might help shape our efforts toward that goal, but they should not be given the power to nullify our moral obligation or to lessen the vigor with which we pursue the goal.

Revising Large-Scale Mathematics Assessment

Some large-scale tests are moving toward a closer alignment with the mathematics curriculum envisioned in the recommendations for reform. For example, since Fall 1990 the Scholastic Aptitude Test is allowing the use of scientific calculators on its Mathematics subtest, and a calculator is now a school-district option on the Iowa Test of Basic Skills which provides users with norms both with and without calculators. The College Entrance Examination Board has even announced its plans to require calculators with graphing capabilities on the AP Calculus Examination by 1994. In efforts to align more closely with the NCTM standards, the Iowa Test of Basic Skills will include a subtest on Estimation with items like those reported in Schoen, Blume, and Hoover (1990), and its Problem Solving subtest will include a number of items aimed at measuring problem-solving processes, not just final answers, as described by Schoen and Oehmke (1980). Figure 1 shows a sample estimation and problem-solving process item. Because these tests continue to use a multiple-choice format, they will not satisfy critics concerned about the narrowing effects of that format on instruction and learning.

1. 4 1/2 + 13 4/5 is between

   a) 18 and 19
   b) 17 and 18
   c) 16 and 17
   d) 19 and 20

2. The school cafeteria had 230 kg of milk to be shared by 46 children. The cook wanted to know how many glasses of milk each child could have. the cook could solve the problem if he also knew:

   a) There are 1000 grams in a kilogram.
   b) Each glass holds 2 kg of milk.
   c) The children all like milk.
   d) Each glass is 8 cm high.

Figure 1. Sample of the types of estimation (#1) and problem-solving process (#2) items to be included on the Iowa Test of Basic Skills.

There are also many current efforts by researchers and measurement specialists to broaden the assumptions and goals of testing to bring it more in line with various aspects of teaching or learning. These efforts include connecting assessment to cognitive development and the cyclical nature of learning (Romberg, Zarinnia, & Williams, 1990), measuring students' pertinent knowledge at the beginning of an instructional sequence and supplementing that with measures of how readily they can understand new skills or procedures just beyond their competence level (Campione, Brown, & Connell, 1988), and assessment of students' schematic organization of knowledge (Marshall, 1988). All of these approaches deserve and require more research and development.

Many practitioners in education have despaired of efforts to shore up familiar and well-articulated forms of testing. They believe that what is needed is an entirely new or reclaimed view of quite different modes of assessment similar to some of those described earlier in this paper for classroom teachers. These modes include observation and analysis of students' work or of students' performance on complex tasks or of portfolios of their work (e.g. Romberg, et al., 1990; Silver & Kilpatrick, 1988; Wolf, et al., 1991). National assessments in New Zealand and Great Britain and state assessments in California, Connecticut, and Vermont have moved in this direction, and many other states and school districts are beginning to follow suit (Wolf et al., 1991).

Even the U.S. national assessment system under development by the New Standards Project, under the direction of Lauren Resnick at the University of Pittsburgh's Learning Research and Development Center, aims to make such performance-based measurement its cornerstone (as reported in the *Report on Education Research*, September 18, 1991). In the same report, Eva Baker, director of the National Center for Research on Evaluation, Standards and Student Testing, warned that:

> Our experiences lead us to believe that valid performance-based assessments can be developed. They take time, conceptual models, and careful empirical work. We fear that the present policy press for such measures will short-cut the process and... result in tasks whose validity cannot be supported. (p. 5)

Performance-based assessment has not had the 50 or so years that the traditional system of educational testing has had to develop its scientific base; therefore, it has less exactness and elegance. That state of affairs should suggest caution to the proponents of large-scale performance-based assessment. On further consideration, however, one might ask why it is important that we have a scientifically, elegant educational measurement system? Is such a system, in itself, an end worth striving for, or is it more appropriately a means to a greater end, namely, supporting and improving the American educational system?

If it is the former, we should probably stick to our traditional testing system. It is unlikely that performance-based assessment, portfolios, and the other open-ended attempts to assess human performance while maintaining much of its complexity will ever be as reliable as a multiple-choice, standardized test. If we see, however, our educational measurement system as a means to support and improve the American educational system, then we should look beyond the goal of scientifically elegant measurement. We should also not forget that, according to the Second International Mathematics Study, which is based on a traditional mathematics achievement test, students in at least eight or ten countries are learning more mathematics than our students are (McKnight, et al., 1987). None of the educational systems in these countries has a testing system that is as efficient and technologically sophisticated as ours.

Furthermore, America's higher education system has, at least until recently, been considered to be the most successful in the world. In Japan, for example, our public schools are scoffed at, but our system of higher education is envied and imitated (Taylor, 1983). It may or may not be a coincidence that, again until recently, higher education has not been dominated and judged by standardized tests; the argument that its mission and goals are too complex to be reduced to multiple-choice tests and quantitative comparisons has prevailed. Perhaps the argument about the complexity of the higher education enterprise applies equally to education at all levels. At any rate, our goal of maintaining a scientifically elegant system of educational measurement for elementary and secondary education appears over the years to have been raised to a level well out of proportion to its potential for supporting and improving our educational system.

The concept of a *test-driven* or *assessment-driven* curriculum seems to be a classic example of giving too much power to measurement and too little to curriculum and instruction. No matter how complex or realistic an assessment procedure is, it seems to have a narrowing effect on teachers and students who are to be judged by it. Their time and effort begins to go toward figuring out what it takes to succeed on the test, and they work toward those skills whether or not the skills are appropriate or valuable mathematics. A good example of this is the mathematical Tripos in Great Britain around 1900. Well-known mathematicians like G. H. Hardy, Bertrand Russell, and J. E. Littlewood took one or two years out of their mathematical educations to work with special tutors to learn test-taking tricks and problem types in preparation for this highly competitive and mathematically sophisticated examination. Winners, including Littlewood, were called wranglers and gained a great deal of public adulation. The Tripos was a mathematically sophisticated, open-ended examination which was considered to be a measurement of higher-order thinking. Yet, Hardy (who finished third on the Tripos) later wrote bitterly about the terrible waste for young British mathematicians who spent years on learning otherwise useless test-taking techniques when they should have been doing mathematical research (Kanigel, 1991).

Like the Tripos examination's domination of mathematics, the assessment-driven curriculum concept reverses the roles of curriculum and assessment. To revitalize the mathematics curriculum, it is necessary that assessment be aligned with the curriculum. Simply developing and using assessment techniques or instruments that are aligned with the new curriculum goals will not guarantee, however, that the curriculum will become a reality. Worse yet, an educational policy which assumes that assessment drives curriculum is likely to divert resources to assessment that should be going to much more important areas of education like classroom teaching, student learning, curriculum development, and teacher education.

Summary

Assessment can and should be a major part of mathematics classroom instruction, but teachers have not usually been well prepared in this area. They should learn a variety of assessment techniques that include but go beyond paper-and-pencil testing and combine that with a solid pedagogical knowledge of mathematics and of the students they teach. Both teacher education and the way teachers teach in their own classrooms must change radically if this ambitious goal is to be reached. Such changes will not come easily, but the effort could result in more professional, reflective mathematics teachers. In turn, such teachers are likely to foster improved student performance in mathematics.

Large-scale achievement tests, research suggests, have been misused in various ways resulting in a number of negative effects on students, teachers, and the curriculum. In an attempt to circumvent some of these negative effects and to better align assessment with current curriculum goals, new or reclaimed forms of large-scale performance-based assessment are now being used in various districts, states, and countries including the New Standards Project in the United States.

The jury is out on the scientific characteristics of these approaches to assessment and on the effects they may have on teachers and students. Whatever the verdict, however, mathematics is in a state of flux. To be well educated mathematically does not mean the same thing as it did 50 years ago or even ten years ago. The rapid changes occurring everywhere in society, especially in technological developments, suggest a need for a thorough and continuous rethinking of our definition of mathematics achievement as operationalized by our traditional tests. Leaders in the mathematics and mathematics education professions believe they should play an important role in helping to shape the mathematical content of, and the nature of mathematics that is implicit in, large-scale assessment techniques.

References

California Mathematics Council. (1989). *Assessment alternatives in mathematics*. Berkeley, CA: EQUALS.

Campione, J. C., Brown, A. L., & Connell, M. L. (1988). Metacognition: On the importance of understanding what you are doing. In R. I. Charles & E. A. Silver (Eds.), *The teaching and assessing of mathematical problem solving* (pp. 93-114). Reston, VA: National Council of Teachers of Mathematics.

Coxford, A. F. (Ed.). (1988). *The ideas of algebra, K-12*. Reston, VA: National of Teachers of Mathematics.

Clarke, D. J., Clarke, D. M., & Lovitt, C. J. (1990). In T. J. Cooney (Ed.), *Teaching and learning mathematics in the 1990s* (pp. 118-129). Reston, VA: National Council of Teachers of Mathematics.

Herscovics, N., & Kieran, C. (1980). Constructing meaning for the concept of equation. *Mathematics Teacher, 73*(8), 572-580.

Hiebert, J., & Behr, M. (Eds.). (1988). *Number concepts and operations in the middle grades*. Reston, VA: Lawrence Erlbaum Associates & National Council of Teachers of Mathematics.

Kanigel, R. (1991). *The man who knew infinity: A life of the genius Ramamanujan*. New York: Charles Scribner's Sons.

Kulm, G. (1990). *Assessing higher order thinking in mathematics*. Washington: American Association for the Advancement of Science.

Leinhardt, G., Zaslavsky, O., & Stein, M. K. (1990). Functions, graphs, and graphing: Tasks, learning, and teaching. *Review of Educational Research, 60*(1), 1-64.

Lester, F. K., & Kroll, D. L. (1990). Assessing student growth in mathematical problem solving. In G. Kulm (Ed.), *Assessing higher order thinking in mathematics* (pp. 53-70). Washington: American Association for the Advancement of Science.

Lester, F. K., & Kroll, D. L. (1991). Evaluation: A new vision. *Mathematics Teacher, 84*(4), 276-284.

Marshall, S. P. (1988). Assessing problem solving: A short-term remedy and a long-term solution. In R. I. Charles & E. A. Silver (Eds.), *The teaching and assessing of mathematical problem solving* (pp. 159-177). Reston, VA: National Council of Teachers of Mathematics.

McKnight, C. C., Crosswhite, F. J., Dossey, J. A., Kifer, E., Swafford, J. O., Travers, K. J., & Cooney, T. J. (1987). *The underachieving curriculum: Assessing U. S. school mathematics from an international perspective*. Champaign, IL: Stipes Publishing Company.

National Advisory Committee on Mathematical Education. (1975). *Overview and analysis of school mathematics grades K-12*. Washington: Conference Board of the Mathematical Sciences.

National Commission on Testing and Public Policy. (1990). *From gatekeeper to gateway: Transforming testing in America*. Chestnut Hill, MA: National Commission on Testing and Public Policy.

National Council of Teachers of Mathematics. (1989). *Curriculum and evaluation standards for school mathematics*. Reston, VA: National Council of Teachers of Mathematics.

National Council of Teachers of Mathematics. (1991). *Professional standards for teaching mathematics*. Reston, VA: National Council of Teachers of Mathematics.

National Research Council. (1989). *Everybody counts: A report to the nation on the future of mathematics education*. Washington: National Academy Press.

National Research Council. (1991a). *Moving beyond myths: Revitalizing undergraduate mathematics*. Washington: National Academy Press.

National Research Council. (1991b). *Counting on you: Actions supporting mathematics teaching standards*. Washington: National Academy Press.

Oakes, J. (1990). *Multiplying inequalities: The effects of race, social class, and tracking on opportunities to learn mathematics and science*. Santa Monica, CA: The RAND Corporation.

Paulos, J. A. (1988). *Innumeracy: Mathematical illiteracy and its consequences*. New York: Hill and Wang.

Popham, W. J. (1987). The merits of measurement-driven instruction. *Phi Delta Kappan, 68*, 679-682.

Romberg, T. A., Zarinnia, E. A., & Collis, K. F. (1990). A new world view of assessment in mathematics. In G. Kulm (Ed.), *Assessing higher order thinking in mathematics* (pp. 21-38). Washington: American Association for the Advancement of Science.

Romberg, T. A., Zarinnia, E. A., & Williams, S. R. (1989). *The influence of mandated testing on mathematics instruction: Grade 8 teachers' perceptions*. Madison, WI: National Center for Research in Mathematical Sciences Education.

Schon, D. A. (1990). *Educating the reflective practitioner*. San Francisco: Jossey-Bass Publishers.

Schoen, H. L., Blume, G., & Hoover, H. D. (1990). Outcomes and processes on estimation test items in different formats. *Journal for Research in Mathematics Education, 21*, 61-73.

Schoen, H. L., & Oehmke, T. (1980). A new approach to the measurement of problem-solving skills. In S. Krulik (Ed.), *Problem solving in school mathematics* (pp. 216-227). Reston, VA: National Council of Teachers of Mathematics.

Silver, E. A., & Kilpatrick, J. (1988). Testing mathematical problem solving. In R. I. Charles & E. A. Silver (Eds.), *The teaching and assessing of mathematical problem solving* (pp. 178-186). Reston, VA: National Council of Teachers of Mathematics.

Smith, M. L. (1991). Put to the test: The effects of external testing on teachers. *Educational Researcher, 20*(5), 8-11.

Stevenson, H. W., Lummis, M., Lee, S. Y., & Stigler, J. W. (1990). *Making the grade in mathematics: Elementary school mathematics in the United States, Taiwan, and Japan*. Reston, VA: National Council of Teachers of Mathematics.

Stigler, J. W., Lee, S. Y., & Stevenson, H. W. (1990). *Mathematical knowledge: Mathematical knowledge of Japanese, Chinese, and American elementary school children*. Reston, VA: National Council of Teachers of Mathematics.

Trafton, P. R. (Ed.). (1989). *New directions for elementary school mathematics*. Reston, VA: National Council of Teachers of Mathematics.

Wagner, S., & Kieran, C. (Eds.). (1989). *Research issues in the learning and teaching of algebra*. Reston, VA: Lawrence Erlbaum Associates & National Council of Teachers of Mathematics.

Webb, N., & Briars, D. (1990). Assessment in mathematics classrooms, K-8. In T. J. Cooney (Ed.), *Teaching and learning mathematics in the 1990s* (pp. 108-117). Reston, VA: National Council of Teachers of Mathematics.

Wolf, D., Bixby, J., Glenn III, J., & Gardner, H. (1991). To use their minds well: Investigating new forms of student assessment. In G. Grant (Ed.), *Review of research in education-volume 17* (pp. 31-74). Washington: American Educational Research Association.

About the author

Harold "Hal" L. Schoen is a Professor of Mathematics Education at the University of Iowa. He is co-director of two NSF-funded projects, the Core-Plus Mathematics Project and the Iowa Assessment Project. His major role in both projects is to develop assessment materials and procedures for the high school mathematics classroom.

OERI

This bulletin was funded by the Office of Educational Research and Improvement, U.S. Department of Education under contract no. RI-88062006. Opinions expressed in this bulletin do not necessarily reflect the positions or policies of OERI or the Department of Education.

ERIC

Educational Resources Information Center. The nationwide information system initiated in 1966 by the U.S. Department of Education. ERIC is the largest and most frequently used database in the world.

SE 053 466

➤ This bulletin is in the public domain and may be freely reproduced ➤